### Computational miRNA target prediction

Following GenMiR++ [2]'s approach, we can combine expression profiles with putative predictions (from sequence analysis).

<table>
<thead>
<tr>
<th>miRNA/mRNA expression</th>
<th>noisy beliefs</th>
<th>Predicted interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Using expression of miRNAs to explain down-regulation of the mRNAs.

Expression of a mRNA $x$ is regressed from the expressions of all miRNAs $y_j$:

$$ x \sim \mathcal{N}(\mu - \sum_j \alpha_j y_j, \sigma^2 I) $$

It is known from experiments that sets of miRNAs can target specific groups of mRNAs.

How can we discover this group structures as well as the interaction networks?

### Existing works

- Class based model: (Mixed membership) Stochastic blockmodels.
- Feature based model: Indian Buffet Process, distance dependent CRP.
- Matrix Factorization.

### Our approach: feature based model, augmenting the IBP

Each entity has $k$ features which can be represented by a binary $n \times k$ matrix $Z$. Therefore, entity $i$ and entity $j$ interact through all groups $k$ such that $z_{ik} z_{jk} = 1$.

**Indian Buffet Process** [1]: a nonparametric Bayesian prior for $Z$.

- The first customer tries Poisson($\alpha$) dishes.
- The $r$th customer tries a previously sampled dish $k$ with probability $\frac{\alpha_k}{Z}$ and samples a Poisson($\alpha$) number of new dishes.
- Customers select dishes independently.

### Introducing dependency between entities

- **Dependency**: A $N \times N$ symmetric prior matrix $W$: $w_{ij}$ indicates the degree that we believe that entity $i$ and $j$ interact ($w_{ij} > 0$ if entities $i$ and $j$ are more likely to interact.)
- For our model, we add the new pairwise potentials on memberships of entities:

$$ P(z_k, \pi_k|\alpha) = \frac{1}{Z!^{K_k-1}} \prod_{k=1}^{K} \frac{\alpha^K_k}{K_k!} \prod_{k=1}^{K} \Phi_{z_k} \exp \left( \sum_{i} \left( (1-z_{ik}) \log(1-\pi_k) + z_{ik} \log \pi_k \right) + \frac{\alpha}{K} \log \pi_k \right) $$

- The probability of a left-ordered equivalence class of binary matrices, $|Z|$ is:

$$ \lim_{K \to \infty} P(|Z|) = \frac{\alpha^{K_k} K_k!}{\prod_{k=1}^{K} \prod_{k=1}^{K} \pi_k} \exp \left( -\alpha \Psi \right) $$

$\Psi = \sum_{k=1}^{K} \frac{N - m_k}{m_k - 1} \frac{m_k!}{N!} \log \left( \frac{m_k!}{N!} \right) $ with $\alpha = \frac{N - m_k}{m_k - 1} \frac{m_k!}{N!}$.

---

### Results

- **Accuracy**
- **Precision**
- **Recall**
- **F1 Score**

- **Simulation**: our method is called GroupMiR

- Our sampling algorithm involves a mix of Gibbs and Metropolis-Hastings steps in C++. 500 samples discarded as burn-in.

- 9 synthetic datasets: 20 miRNAs and 200 mRNAs, $K = 5$. Noise was introduced to the true binary interaction matrix with different probabilities.

### MicroRNA Networks in Mouse Lung Organogenesis

We performed Gene Ontology (GO) enrichment analysis on identified clusters. Several cell division categories are enriched in cluster (b) (expected when dealing with a developing organ). Other significant functions include organelle organization and apoptosis which also are associated with development (cluster (c)).

Cluster (a) includes 2 members of the miR-17-92 set, which is known to be critical to lung organogenesis. MiRNA families miR-30, miR-29, miR-20 and miR-16, all identified by our method, were also reported to play roles in the early stages of lung organogenesis.

It is important to point out that we did not filter miRNAs explicitly based on expression. However, most miRNA identified by our method are differentially expressed since the method relied on their expression to explain the observed mRNA expression.

---

Hai-Son Le, Ziv Bar-Joseph  
Machine Learning Department, Carnegie Mellon University

{-hple,zivbj}@cs.cmu.edu

Systems Biology Group (http://www.sb.cmu.edu)